

## The Rat Casein Multigene Family

FINE STRUCTURE AND EVOLUTION OF THE  $\beta$ -CASEIN GENE\*

(Received for publication, October 19, 1984)

William K. Jones, Li-yuan Yu-Lee†, Shirley M. Clift, Terry L. Brown§, and Jeffrey M. Rosen¶

From the Department of Cell Biology, Baylor College of Medicine, Houston, Texas 77030

Eight overlapping phage clones, spanning 34.4 kilobase pairs of genomic DNA, containing the 7.2-kilobase pair rat  $\beta$ -casein gene have been isolated and characterized. The first 510 base pairs (bp) of 5' flanking, 110 bp of 3' flanking, and all the exon/intron junctions have been sequenced. The  $\beta$ -casein gene contains 9 exons ranging in size from 21 to 525 bp. We have attempted to identify potential regulatory elements by searching for regions of sequence homology shared between milk protein genes which respond similarly to lactogenic hormones and by searching for previously reported hormone receptor-binding sites. Within the conserved first 200 bp of 5' flanking sequences 3 regions of greater than 70% homology were observed between the rat  $\beta$ - and  $\gamma$ -casein genes. One of these contains a region 90% homologous to the chicken progesterone receptor-binding site. The conserved 5' noncoding region, the highly conserved signal peptide, and the hydrophobic carboxyl-terminal region of the protein are each encoded by a separate exon. In contrast the evolutionarily conserved phosphorylation site of  $\beta$ -casein is formed by an RNA-splicing event. The exons which encode the phosphorylation sites of  $\beta$ -casein appear to have resulted from an intragenic duplication. Based upon the exon structure of the casein genes, an evolutionary model of intragenic and intergenic exon duplications for this gene family is proposed.

The  $\beta$ -casein gene is a member of a small gene family, containing at least five genes, responsive to both steroid and peptide hormones (1). Although these genes are each induced during lactation, each casein mRNA exhibits different kinetics of induction and extents of accumulation due to complex multihormonal regulation of transcription and mRNA stability (2, 3). In the rat  $\beta$ -casein mRNA accounts for  $\approx$ 20% of the cellular mRNA during lactation, thus encoding the predominant calcium-sensitive casein in this species (4). Definition of the structure of the casein genes, their gene products, and evolutionary interrelationships is a necessary prerequisite

for elucidating the respective mechanisms by which both peptide and steroid hormones act. Analysis of rat, bovine and guinea pig casein cDNA sequences has demonstrated considerable divergence among the individual members of the casein gene family (5-8). In fact, the casein proteins are among the most rapidly diverging protein families studied (9). Only three regions of the casein mRNAs are conserved: the 5' noncoding region, the signal peptide-coding regions, and the regions encoding the sites of phosphorylation and calcium binding (8). The casein gene family is believed to have evolved by both intragenic and intergenic duplication of a primordial gene containing a phosphorylation and calcium-binding site and a signal peptide sequence (8). This hypothesis is supported by genetic data which revealed the clustering of the bovine casein genes and the localization of the mouse  $\alpha$ -,  $\beta$ - and  $\gamma$ -casein genes to chromosome 5 (10-12).

We have previously reported the characterization of the large and complex 15-kb<sup>1</sup>  $\gamma$ -casein gene containing at least 9 exons and an unusual Goldberg-Hogness sequence (TTTAAAT (13)). We report the isolation, the fine mapping, and the sequencing of all the exons, the exon/intron junctions, and the 5' and 3' flanking regions of the  $\beta$ -casein gene. In addition, the analysis of an apparent primary gene transcript as well as various processed forms of this gene product are reported. From the comparison of the  $\gamma$ -casein and  $\beta$ -casein 5' flanking regions, three sets of conserved sequences are described. Finally, a model of the evolution of this gene family, based upon the exon structure of the casein genes, is proposed.

EXPERIMENTAL PROCEDURES<sup>2</sup>

## RESULTS AND DISCUSSION

As reported in the Miniprint Section, 8 overlapping phage clones, spanning 34.4 kb of genomic DNA, containing the  $\beta$ -casein gene locus have been isolated and mapped. The orientation of the gene was determined by hybridization of various restriction enzyme digestions to specific 5' and 3' cDNA probes. The comparison of restriction enzyme digestions of total rat DNA with the isolated clones showed that the latter were not rearranged. In addition to the 7.2-kb  $\beta$ -casein gene, these clones contain 14.6 kb of 5' flanking DNA and 12.6 kb

\* This paper is the third in a series. The first is "The Rat Casein Multigene Family. I. Fine Structure of the  $\gamma$ -Casein Gene" (13). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† Recipient of American Cancer Society Grant BC 425.

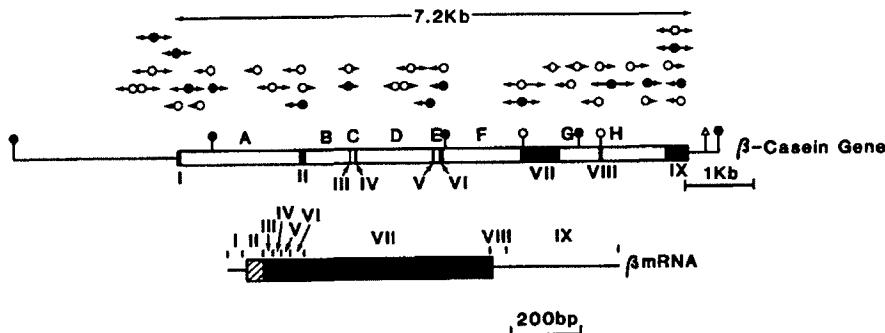
§ Recipient of National Institutes of Health Postdoctoral Fellowship GM09392.

¶ Supported by National Institutes of Health Grant CA16303. To whom reprint requests should be addressed: Department of Cell Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030.

<sup>1</sup> The abbreviations used are: kb, kilobase pair; bp, base pair(s); 1  $\times$  SSC, 0.15 M NaCl, 0.015 M Na citrate; BSA, bovine serum albumin; SDS, sodium dodecyl sulfate.

<sup>2</sup> Portions of this paper (including "Experimental Procedures," part of "Results," Figs. 1-5, and Table I) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are available from the Journal of Biological Chemistry, 9650 Rockville Pike, Bethesda, MD 20814. Request Document No. 84M-3529, cite the authors, and include a check or money order for \$4.00 per set of photocopies. Full size photocopies are also included in the microfilm edition of the Journal that is available from Waverly Press.

FIG. 6. Sequencing strategy of the  $\beta$ -casein gene. A map of the  $\beta$ -casein gene is shown (middle). Closed boxes represent exons (I-IX) and open boxes represent introns (A-H). Each arrow above represents one sequencing gel. Open circles represent 3' labeling, and closed circles represent 5' labeling. The region of the  $\beta$ -mRNA which each exon encodes is indicated at the bottom of the figure.



of 3' flanking DNA, none of which has yet been shown to overlap another functional casein gene. The small size of the exons made their preliminary localization by DNA blots particularly difficult. Thus, alternative methods such as reduced stringency RNA sandwich blots and R-loop analysis had to be employed. The detection of apparently full-length 7.5-kb RNA transcripts confirmed the predicted size of the  $\beta$ -casein gene. These results are described in detail in the Miniprint Section.

**Sequencing of the  $\beta$ -Casein Gene**—The structure of the  $\beta$ -casein gene was determined by the sequencing strategy shown in Fig. 6. Six thousand base pairs were sequenced, of which 66% including all the exon/intron boundaries, the first 510 bp of 5' flanking, and 110 bp of 3' flanking sequences were sequenced in both directions (Fig. 7). Only 8 base pair differences were observed in comparison to the published  $\beta$ -cDNA sequences altering 1 amino acid (5). These changes may represent sequencing ambiguities or allelic differences. The 7.2-kb  $\beta$ -casein gene contains 9 exons separated by 8 introns and thus appears similar in organization to the  $\gamma$ -casein gene which contains a minimum of 9 exons (the sizes of all the exons and introns are listed in Table I) (13). All the exon/intron junctions of the coding exons are located between codons, and the exon/intron splice junctions are in agreement with the canonical sequence suggested by Mount (14).

**5' Flanking Sequence Comparison**—Our laboratory and others have been studying the structure of several milk protein genes which are all expressed in the mammary gland in response to the lactogenic hormones prolactin and hydrocortisone. From the comparison of the flanking regions of various milk protein genes, it has been possible to identify regions of shared homology which may represent regulatory regions. The  $\gamma$ - and  $\beta$ -casein 5' flanking regions were initially compared since the expression of both of these evolutionarily related genes is induced by prolactin and glucocorticoids and inhibited by progesterone (15).

The  $\beta$ - and  $\gamma$ -casein gene 5' flanking regions were first compared by dot matrix analysis using the stringent criteria that 7 of 10 nucleotides must be identical to be scored positive (Fig. 8). Three regions of sequence satisfying these criteria were observed within the conserved 5' flanking regions. In the  $\beta$ -casein gene the proximal region of homology lies between -48 and -63, the medial between -106 and -119, and the distal between -130 and -165. In the  $\gamma$ -casein these regions are at the same positions  $\pm$ 3 bp. The proximal regions are composed of  $\approx$ 25% G/C base pairs, and the distal regions have the highest G/C content of  $\approx$ 40% G/C base pairs. Aside from these three regions, the two 5' flanking sequences show poor homology. Using the same criteria, no homology was observed between the first 200 bp of the  $\beta$ - or  $\gamma$ -casein gene 5' flanking regions and the corresponding regions of the whey acidic protein (16) or  $\alpha$ -lactalbumin (17) genes. These genes,

although not evolutionarily related to the caseins, display similar developmental and hormonal regulation (2).

The  $\beta$ - and  $\gamma$ -casein genes possess distinctly different Goldberg-Hogness sequences. The Goldberg-Hogness sequence is believed to play an important role in determining the correct site of transcription initiation (18). In the  $\beta$ -casein gene, the TATA box (TATATAT) shows a greater homology to the canonical Goldberg and Hogness sequence (TATA $\frac{T}{A}$  $\frac{T}{A}$ ) (18)

than the more divergent TATAs (TTTAAAT) of  $\gamma$ -casein,  $\alpha$ -casein,<sup>3</sup> and whey acidic protein genes (13, 16). It has been reported in other genes that a T at the second position in the Goldberg-Hogness sequence, rather than an A, reduces its *in vitro* promoter efficiency (19). This difference in the TATA sequences may be involved in the relatively higher levels of  $\beta$ -casein gene expression in comparison to the other casein genes, but this remains to be established. The  $\beta$ -casein gene does not have an identifiable CAAT box (18) at the usual location -80 bp from the CAP site. The sequence CAAAT is found near -58 bp and within the proximal region of conservation observed within the casein family.

Both strands of sequenced regions of the  $\beta$ -casein gene have been searched for hormonally responsive elements such as the reported binding sites of the progesterone (20-22) and glucocorticoid (23) receptors as well as a sequence common to estrogen-regulated genes (23). The hexanucleotide (TGTTC) common to the reported glucocorticoid receptor-binding sites (23) was found five times in the  $\beta$ -casein gene at -510 in the 5' flanking region, at 26 within exon I, at 2484 within exon III, at 5051 within exon VII, and 5800 within intron G. None of the surrounding sequences showed greater than 80% homology to any of the four glucocorticoid receptor-binding sites reported by Renkawitz *et al.* (23). No sequences identical to the nanomer common to estrogen-regulated genes were observed. No sequences showed greater than 80% homology to the progesterone receptor-binding site reported by Mulvihill (20). However, one sequence between -157 and -143 (TGTCCCCAGAATT) did have 86% homology to the sequence between -184 and -171 of the chicken ovalbumin gene (TGTTACCCAGAATT) which has been reported by Compton *et al.* (21) to be within the progesterone receptor-binding site defined by DNase I footprinting (21). This region has also been shown to be involved in progesterone and estrogen regulation by deletion analysis (22). Interestingly, this is within the distal conserved region of the  $\beta$ - and  $\gamma$ -casein genes.

The three conserved 5' flanking regions of the casein genes are candidates for regulatory elements. The conservation, which is greater than most of the coding regions of  $\beta$ - and  $\gamma$ -casein, indicates that these regions have been selected for

FIG. 7. Sequence of the  $\beta$ -casein gene. The 5' flanking, 3' flanking, and intron sequences are denoted by lower case letters. Exons are represented by upper case letters, and the amino acids encoded are noted where appropriate. The lengths of unsequenced regions are estimated from restriction mapping data.

during evolution. The resemblance of the proximal region to the CAAT sequence and the distal region to the progesterone receptor-binding site suggest possible functions for these two regions which can be tested experimentally. To accomplish this, the entire  $\beta$ -casein gene obtained from the genomic phage clone B13 (Fig. 1) has been inserted into a bovine papilloma virus expression vector to test the possible functional role of these and other sequences.<sup>4</sup>

**3' Flanking Sequence Comparison**—The first 110 bp of the 3' flanking region of the  $\beta$ -casein gene have been sequenced in both directions. As previously reported, the  $\beta$ -casein gene has a canonical poly(A) sequence (AATAAA) located 16 bp 5' to the polyadenylation site which for the  $\beta$ -casein gene is TA (5). Four bp 3' to the A nucleotide is the sequence CAGTTG, a close derivative of the pentanucleotide sequence CAYUG reported by Berget (24). This pentanucleotide is proposed to hybridize with U4 RNA and be involved in mRNA

<sup>4</sup> Y. David-Inouye, personal communication.

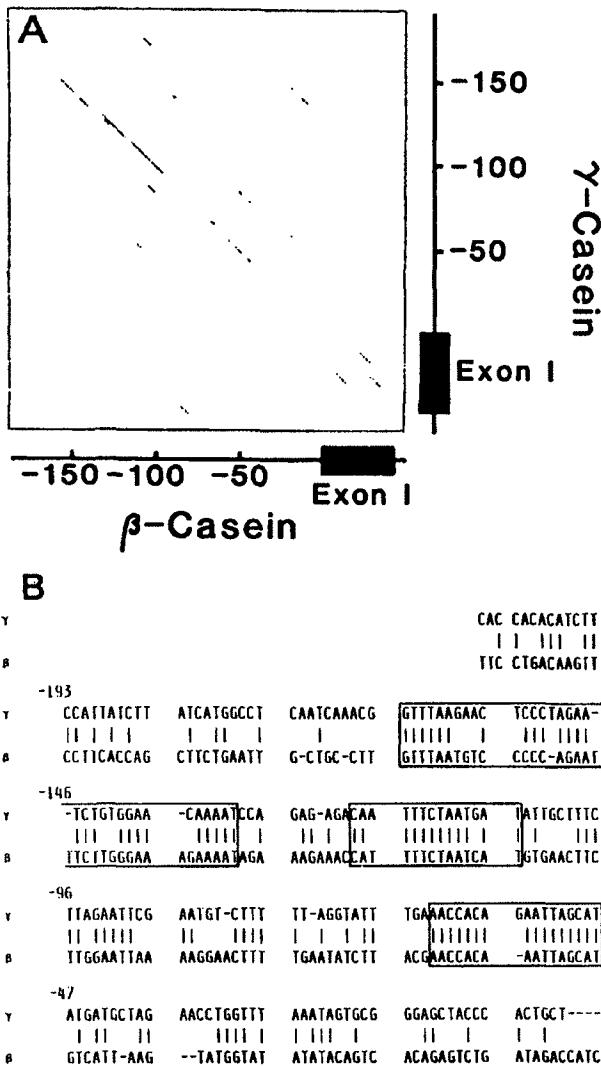


FIG. 8. Panel A, dot matrix analysis of the  $\beta$ - and  $\gamma$ -casein 5' regions. The first 200 bp of each gene were compared using a dot matrix analysis. The location of the  $\beta$ - and  $\gamma$ -casein gene sequences are indicated by the maps along the X and Y axis. To be scored as positive, at least 7 out of 10 nucleotides were required to match. Panel B, comparison of the first 200 bp of 5' flanking DNA. The sequences of the  $\beta$ - and  $\gamma$ -casein 5' flanking DNA were aligned to give maximum homology as indicated by the dot matrix analysis. The sequence is arranged 5' to 3' with the -1 nucleotide in the lower right-hand corner. The Goldberg-Hogness sequences and the three conserved regions are enclosed by boxes.

cleavage site selection. The downstream location of the pentanucleotide and the lack of such a sequence between the AATAAA and the cleavage site indicate that the  $\beta$ -casein gene belongs to the class II SV40 late type of polyadenylation sites.

**Casein Gene Structure**—In addition to identifying potential regulatory elements it was also of interest to determine possible evolutionary relationships among the casein genes. The casein proteins are under unusual evolutionary pressure, because unlike enzymes they do not need to maintain the stringent geometry of an active site but are still required for the reproductive success of mammals. Thus the genes have undergone rapid divergence while still encoding a functional protein. Gilbert (25) and Blake (26) have proposed that proteins are composed of functional domains, and these domains are encoded by an exon or a group of exons. Thus, new

proteins could be easily assembled from previously evolved functional domains by the recruitment of the corresponding exons to form a new gene. It was intriguing to apply this model to the analysis of the casein genes. To fulfill their nutritional role the caseins must perform three functions: 1) to be secreted; 2) to form protein aggregates termed micelles; and 3) to be phosphorylated to allow  $\text{Ca}^{2+}$  binding and transport. When the exon structure of the  $\beta$ -casein gene was examined, the first two functions are clearly carried out by polypeptide sections encoded by individual exons.

The conserved 5' noncoding region and the signal peptide which allows casein protein secretion are encoded by exons I and II, respectively (Figs. 6 and 7). Exon I contains most of the 5' noncoding sequence of the gene. The site of transcription initiation is assumed to start at the A residue identified from primer extension experiments (5). This exon is similar in size as well as in sequence to the first exon of the  $\gamma$ - and the rat  $\alpha$ -casein genes.<sup>3</sup> Exon II encodes the remainder of the 5' noncoding sequence, the entire signal peptide sequence, and the first two amino acids of the mature protein. The signal peptides are the most conserved region of the caseins and presumably constitute a functional domain.

Casein proteins aggregate and form micelles by the interaction of their carboxyl-terminal hydrophobic domains (27). Micelle formation allows milk to contain higher concentrations of protein and calcium phosphate complexes than would be possible for individual protein molecules. The hydrophobic domains do not need to maintain a constrained geometry since they do not appear to close pack. This model of a loose casein micelle structure is supported by the absence of x-ray diffraction spacing indicating that the caseins do not form an ordered structure (28). The reported density of micelles formed from equal weights of  $\alpha$ - and  $\beta$ -bovine caseins is  $\approx 0.3$ – $0.4$  g/cc (27) in contrast to the significantly higher density of  $\approx 1.4$  g/cc for a close packed globular protein (29), which also supports the loose packing model. This loose packing allows caseins to accommodate more changes in their amino acid composition than can a typical protein. Thus it is not surprising that the coding regions of the casein genes appear to have diverged rapidly. The  $\beta$ -casein hydrophobic domain, except for the terminal amino acid, is encoded by exon VII. This region is composed of 44% hydrophobic residues as compared with the more hydrophilic amino-terminal region of the protein which contains 26% hydrophobic residues. Thus, the second casein function, i.e. micelle formation, is carried out by a region of the protein encoded by a single exon. However, since the exon(s) which encode the hydrophobic domains of the other rat casein genes have not been

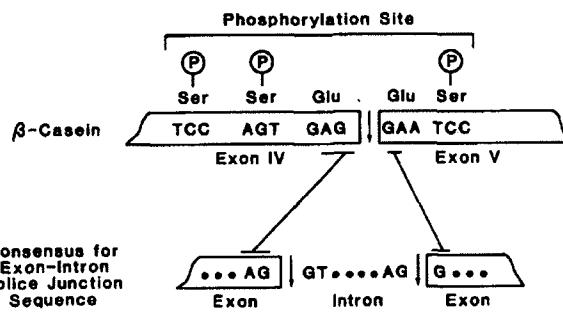


FIG. 9. Structure of the  $\beta$ -casein major phosphorylation site. The exon structure of the  $\beta$ -casein major phosphorylation site is shown above. Exon IV encodes the Ser-Ser-Glu residues, and exon V the Glu residue. The relationship of the conserved GAG and GAA glutamate codons to the consensus exon/intron splice junction is illustrated below.

sequenced, it is not known if they also have a similar exon structure.

The relationship of the exons encoding the phosphorylation sites to the structure of this conserved functional domain is more complex. The sites of phosphorylation and calcium binding are conserved among the casein genes at both the amino acid and nucleic acid levels (8). Phosphorylation sites are typically found in the amino-terminal region of the casein proteins. The casein kinase phosphorylates a serine two residues to the amino-terminal side of the sequence, Ser-X-Y, where Y is either a phosphorylated serine or an acidic residue, usually glutamic acid (30). Minor phosphorylation sites contain only one glutamic acid residue, Ser-X-Glu, while the more common major phosphorylation sites have two glutamic acid residues, Ser-Ser-Glu-Glu. The coding region for this series of residues, one of the most highly conserved portions of the casein mRNA, is as follows (8).

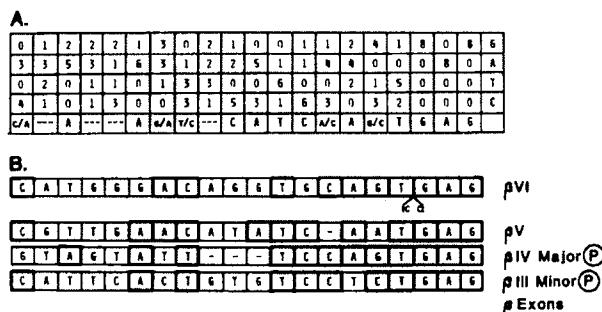


FIG. 10. Conservation of casein phosphorylation sites. The frequency of each base at a given position within the major phosphorylation site of the rat  $\alpha$ - and  $\gamma$ -caseins is shown in panel A. The conserved GAG at right corresponds to the amino-terminal glutamate codon of the major phosphorylation site. The bottom row of panel A records the most frequently used nucleotide at that position. In panel B, the nucleotide sequences of the 3' ends of exons III-VI of rat  $\beta$ -casein are listed. When a particular nucleotide agrees with the most frequently used nucleotide from panel A, it is enclosed by a bold box.

TCN-AGN-GAG-GAA  
Ser-Ser-Glu-Glu

The conserved major phosphorylation site of  $\beta$ -casein is not encoded by a single exon but rather formed by an RNA-splicing event (Fig. 9). The conserved sequence of the major phosphorylation site is split with the Ser-Ser-Glu residues, the equivalent of a minor phosphorylation site, encoded by the 3' region of exon IV and the final glutamic acid residue encoded by the 5' end of exon V.

How do these exons relate to the exons encoding functional domains postulated by Gilbert (25) and Blake (26)? It appears that in the caseins the "functional phosphorylation domain," corresponding to an exon, is a minor phosphorylation site. These minor sites are converted to major sites by a splicing event which juxtaposes a glutamate codon with the minor phosphorylation site to form a major phosphorylation site.

Interestingly, those exons which have junctions between codons, such as the caseins, and which conform to the consensus sequences should frequently code for Glu-Glu residues at their exon/intron junctions (14). The 3' exon sequence C (AG | GT) A (AGT) could only code for glutamine, lysine, and glutamate while the codon starting with the 5' exon sequence (AG | G) would on average encode glutamate one-fourth of the time. Thus, it appears that Glu-Glu residues, characteristic of the major phosphorylation site, would be commonly encoded by exons which meet the above criteria.

The positioning of introns within the phosphorylation coding region may explain the complete conservation of the glutamate codons of the phosphorylation site. The amino-terminal glutamate codon is always GAG while that of the carboxyl-terminal is GAA. The position of these codons at the 3' and 5' ends of exons, respectively, means that they form part of the exon/intron consensus sequence reported by Mount (14) (Fig. 9). Thus the codons are under selective pressure not only to maintain the phosphorylation site but also to maintain correct splicing.

The split architecture of the conserved phosphorylation site

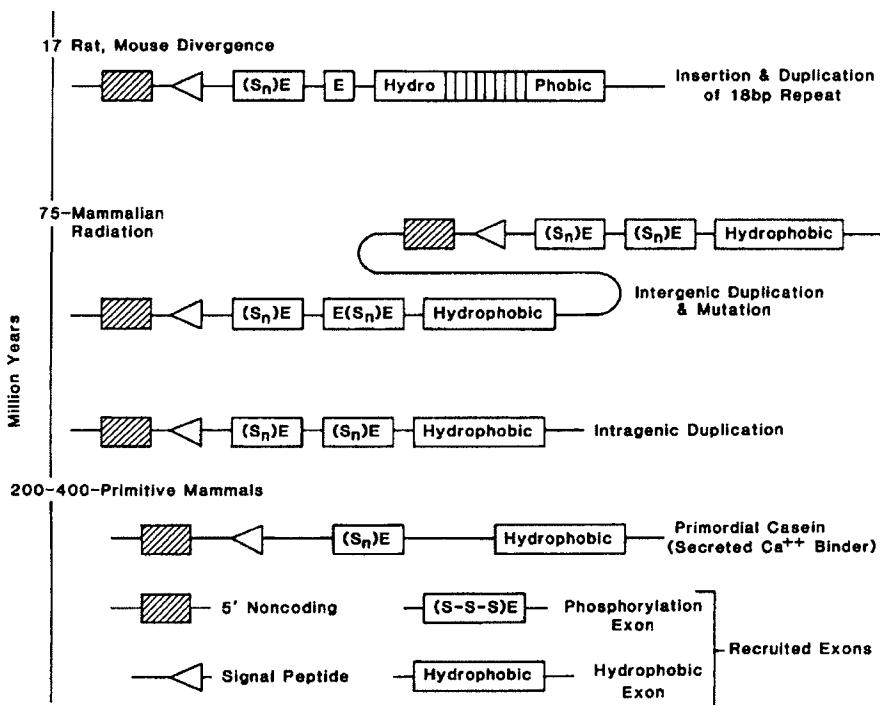


FIG. 11. Model of the evolution of the casein gene family. For discussion of the model, see text. For estimates of the rates of divergence, see Hobbs and Rosen (8). E, glutamic acid; S, serine.

may reveal a relationship between the exon structure of the  $\beta$ -casein gene and the structure of the  $\beta$ -casein protein. Craik (31) has hypothesized that exon/intron junctions often encode peptides found at the surface of the native protein. If this were true for the  $\beta$ -casein protein, the phosphorylation sites would exist at its surface. A surface location for the phosphorylation site would allow the polar residues, serines and glutamic acids, which form the phosphorylation site to be at the protein surface. Also, a surface location would seem advantageous since it would be more accessible to the Golgi casein kinase activity as well as for  $\text{Ca}^{2+}$  binding.

Exons IV and V, as well as exons III and VI, appear to have originated by an exon duplication event. The four exons are quite similar in sequence, especially at their 3' ends, both to each other and to the phosphorylation sites of rat  $\alpha$ - and  $\gamma$ -casein (Fig. 10). It was not surprising that exons III and IV of the  $\beta$ -casein gene were similar to the conserved phosphorylation sites since they each encode a phosphorylation site, but the similarities of the 3' regions of exons V and VI were surprising since they do not encode a functional phosphorylation site. We hypothesize that this homology is the result of an earlier duplication which produced the 4 present-day exons.

A similar duplication seems to have occurred in the bovine  $\alpha_{s1}$ -casein gene (7). Two 24-bp repeats were observed in the cDNA sequence, each of which encode a minor phosphorylation site at its 3' end. If they correspond to exons, the major phosphorylation site of the bovine  $\alpha_{s1}$ -casein would also be formed by a splicing event identical to that in the rat  $\beta$ -casein. Six other 24-nucleotide blocks with 54% homology to the 24-bp repeat noted above were also found, one of which also encodes a minor phosphorylation site at its 3' end. When this latter 24-bp block is compared to the preferred sequences (Fig. 10A), 14 out of the possible 16 nucleotides are identical. If the bovine  $\alpha_{s1}$ -casein gene structure is like the rat  $\beta$ -casein, it is predicted that these blocks will each be encoded by separate exons.

In addition to the sequence homology, the small sizes and the grouping of the  $\beta$ -casein exons III–VI, i.e. two sets of small exons separated by a small intron (III and IV, V and VI), suggest that they are the result of two duplication events. One small exon may have been duplicated to give two small exons separated by a small intron. Later, this pair of exons reduplicated to give the two pairs of exons observed today.

The 3' noncoding region of the  $\beta$ -casein gene, in contrast to the  $\gamma$ -casein gene, is divided into two exons, VIII and IX. Exon VIII encodes the carboxyl-terminal amino acid and the termination codon. However, the rest of the 3' noncoding region, including the polyadenylation signal AATAAA, is contained in exon IX. In contrast, the final exon of  $\gamma$ -casein contains the entire 3' region of the gene starting with the carboxyl-terminal amino acid.<sup>3</sup>

**Casein Gene Evolution**—A hypothetical model of the evolution of the calcium-sensitive casein gene family is proposed in Fig. 11. The casein gene is composed of several distinct regions each encoded by its own exon or exons. We propose that these exons are the result of exon recruitment to form a primitive casein gene followed by intra- and intergenic duplications. Based upon the rates of divergence of the casein signal peptides (8) a calcium-sensitive casein-like gene appears to have originated 300 million years ago at the time of the appearance of primitive mammals. This casein-like gene may have been the result of exon recruitment bringing together the basic elements of the modern caseins: a 5' noncoding region exon, a signal peptide exon, an exon with a minor phosphorylation site at its 3' end, and a hydrophobic domain exon.

Between 300 million years ago and the mammalian radiation, 75 million years ago, two types of duplications appear to have occurred. The phosphorylation exon was duplicated intragenically to create the several small exons observed in the  $\beta$ -casein gene. These exons encode several minor phosphorylation sites which may have been converted to a major phosphorylation site by mutating the first codon of the downstream exon to a glutamate codon. Intergenic duplications also occurred creating the individual members of the casein gene family. It is unlikely that  $\kappa$ -casein, a noncalcium-sensitive casein with a distinct signal peptide sequence, is a product of these intragenic duplications (7). Finally, before the divergence of the rat and the mouse 17 million years ago, the  $\alpha$ -casein gene underwent an insertion of 9½ copies of an 18-bp repeat which is flanked by a direct repeat (CCAA) (8). This accounts for the smaller size of the other mammalian  $\alpha$ -caseins in comparison to the rat and mouse (7).

This model explains the strong homologies between the 5' noncoding regions and the signal peptides of the casein genes. It also accounts for the duplication of the phosphorylation sites even though they are split by an intron. The validity of this model should be tested by determining the exon structure of other casein genes from both the rat and other mammalian species.

**Acknowledgments**—We thank Dr. Miles Mace for his assistance in R-loop analysis, Sara Rupp for technical assistance, Drs. Tom Sargent, Linda Jagodzinski, and James Bonner for providing the rat DNA libraries, Dr. C. Lawrence for his assistance in computer analysis, and Patricia Kettlewell for preparation of this manuscript.

## REFERENCES

1. Topper, Y. S. (1970) *Recent Prog. Horm. Res.* **26**, 286–308
2. Hobbs, A. A., Richards, D. A., Kessler, D. J., and Rosen, J. M. (1982) *J. Biol. Chem.* **257**, 3598–3605
3. Guyette, W. A., Matusik, R. J., and Rosen, J. M. (1979) *Cell* **17**, 1013–1023
4. Rosen, J. M., Woo, S. L. C., and Comstock, J. P. (1975) *Biochemistry* **14**, 2895–2903
5. Blackburn, D. E., Hobbs, A. A., and Rosen, J. M. (1982) *Nucleic Acids Res.* **10**, 2295–2307
6. Hall, L., Laird, J. E., Pascall, J. C., and Craig, R. K. (1984) *Eur. J. Biochem.* **396**, 1–5
7. Stewart, A. F., Willis, I. M., and Mackinlay, A. G. (1984) *Nucleic Acids Res.* **12**, 3895–3907
8. Hobbs, A. A., and Rosen, J. M. (1982) *Nucleic Acids Res.* **10**, 8079–8098
9. Dayhoff, M. O. (1976) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed) Vol. 5, Suppl. 2, pp. 261–262, National Biomedical Research Foundation, Bethesda, MD
10. Matyukov, V. S., and Urnysh, A. P. (1980) *Genetika* **16**, 884–886
11. Grosclaude, F., Mercier, J.-C., and Ribadeau-Dumas, B. (1973) *Neth. Milk Dairy J.* **27**, 328–340
12. Gupta, P., Rosen, J. M., D'Eustachio, P., and Ruddle, F. H. (1982) *J. Cell Biol.* **93**, 199–204
13. Yu-Lee, L., and Rosen, J. M. (1983) *J. Biol. Chem.* **258**, 10794–10804
14. Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472
15. Rosen, J. M., O'Neal, D. L., McHugh, J. E., and Comstock, J. P. (1978) *Biochemistry* **17**, 290–297
16. Campbell, S. M., Rosen, J. M., Hennighausen, L. G., Strehl-Jurk, U., Sippel, A. E. (1984) *Nucleic Acids Res.* **12**, 8685–8697
17. Qasba, P. K., and Safaya, S. K. (1984) *Nature* **308**, 377–380
18. Breathnach, R., and Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383
19. Concino, M., Goldman, R. A., Caruthers, M. H., and Weinmann, R. (1983) *J. Biol. Chem.* **258**, 8493–8496
20. Mulvihill, E. R., LePennec, J.-P., and Chambon, P. (1982) *Cell* **28**, 621–632
21. Compton, J. G., Schrader, W. T., and O'Malley, B. W. (1983) *Proc. Natl. Acad. Sci. U. S. A.* **80**, 16–20
22. Dean, D. C., Gope, R., Knoll, B. J., Riser, M. E., and O'Malley,

B. W. (1984) *J. Biol. Chem.* **259**, 9967-9970

23. Renkawitz, R., Schütz, G., Dietmar, V. D. A., and Beato, M. (1984) *Cell* **37**, 503-510

24. Berget, S. M. (1984) *Nature* **309**, 179-182

25. Gilbert, W. (1978) *Nature* **271**, 501

26. Blake, C. C. F. (1978) *Nature* **273**, 267

27. Waugh, D. F., Creamer, L. K., Slattery, C. W., and Dresdner, G. W. (1970) *Biochemistry* **9**, 786-795

28. Tuckey, S., Roche, H., and Clark, G. L. (1938) *J. Dairy Sci.* **21**, 767-776

29. Metzler, D. E. (1977) in *Biochemistry. The Chemical Reactions of Living Cells*, 1st Ed., p. 75, Academic Press, New York

30. Mercier, J.-C. (1981) *Biochimie* **63**, 1-17

31. Craik, C. S., Rutter, W. J., and Fletterick, R. (1983) *Science* **220**, 1125-1129

32. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning, A Laboratory Manual*, p. 104, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

33. Sargent, T. D., Wu, J.-R., Sala-Trepas, J. M., Wallace, R. B., Reyes, A. A., and Bonner, J. (1979) *Proc. Natl. Acad. Sci. U. S. A.* **76**, 3256-3260

34. Blattner, F. R., Williams, B. G., Bleckl, A. E., Denniston-Thompson, K., Farber, H. E., Furlong, L.-A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumm, J. W., Sheldon, E. L., and Smithies, O. (1977) *Science* **196**, 161-169

35. Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503-517

36. Smith, G. E., and Summers, M. D. (1980) *Anal. Biochem.* **109**, 123-129

37. Johnson, M. L., Levy, J., Supowitz, S. C., Yu-Lee, L., and Rosen, J. M. (1983) *J. Biol. Chem.* **258**, 10805-10811

38. Prentki, P., Karch, F., Iida, S., and Meyer, J. (1981) *Gene (Amst.)* **14**, 289-299

39. Vieira, J., and Messing, J. (1982) *Gene (Amst.)* **19**, 259-268

40. Chaconas, G., and van de Sande, J. H. (1980) *Methods Enzymol.* **65**, 75-85

41. Holmes, D., and Quigley, M. (1981) *Anal. Biochem.* **114**, 193-197

42. Birnboim, H. C., and Doly, J. (1979) *Nucleic Acids Res.* **7**, 1513-1523

43. Rosen, J. M. (1976) *Biochemistry* **15**, 5263-5271

44. Messing, J. (1983) *Methods Enzymol.* **101**, 20-78

45. Maxam, A. M., and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560-564

46. Maizel, J. V., Jr., and Lenk, R. P. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78**, 7655-7669

47. Schibler, U., and Weber, R. (1974) *Anal. Biochem.* **58**, 225-230

48. Chirgwin, J., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. (1979) *Biochemistry* **18**, 5294-5298

49. Bailey, J. M., and Davidson, N. (1976) *Anal. Biochem.* **70**, 75-85

50. Shinnick, T. M., Lund, E., Smithies, O., and Blattner, F. R. (1975) *Nucleic Acids Res.* **2**, 1911-1929

## Supplementary Material to

The Rat Casein Multigene Family.  
Fine Structure and Evolution of the  $\beta$ -Casein GeneWilliam K. Jones, Li-yuan Yu-Lee, Shirley W. Clift  
Terry L. Brown & Jeffrey M. Rosen

## EXPERIMENTAL PROCEDURES

**Materials** - All restriction enzymes were purchased from Bethesda Research Laboratories, New England Biolabs, or Boehringer Mannheim, and used either in the buffers recommended by the suppliers or the three salt buffers recommended by Maniatis et al. (32). DNA polymerase I was from Bioblock DNA polymerase I Klenow fragment was from Bethesda Research Laboratories, or Boehringer Mannheim. The DNA ligase, bacterial alkaline phosphatase and T4 polynucleotide kinase were purchased from Bethesda Research Laboratories. Calf intestinal phosphatase was from Boehringer Mannheim. Elutip-D columns were purchased from Schleicher and Schuell. Type XAR-5 x-ray film was purchased from Eastman Kodak.

**Library Screening and Phage DNA Preparation** - Two rat genomic libraries prepared from random partial Eco RI (33) or  $\lambda$  Sfi I digests of Sprague-Dawley rat DNA and cloned into Charon 4A phage (34) were kindly provided by Drs. T. Sargent, B. Wallace and J. Bonner, and Drs. L. Jagodzinski and J. Bonner, respectively (California Institute of Technology). The libraries were screened and the phage DNA prepared as described previously (13).

**Gel Transfer and Hybridization** - Cloned DNA fragments were transferred to nitrocellulose filters either by the method of Southern (35) or bidirectionally by the method of Smith and Summers (36). Filters were then treated and hybridized as previously described (13). Genomic DNA was prepared from mammary glands and DNA blots performed as described (37).

**Plasmid Subcloning and Plasmid DNA Preparation** - DNA fragments from phage clones were subcloned into either pBR322 (38) or pUC8 (39) plasmids. Total phage DNA was digested with Eco RI and treated with either bacterial alkaline phosphatase (40) or calf intestinal phosphatase (32). Ligations and transformations were done as described previously (13). Plasmid miniprepacations were performed by either the method of Holmes and Quigley (41) or by the alkaline lysis method of Birnboim and Doly (42). Plasmid DNA was prepared as described previously (13).

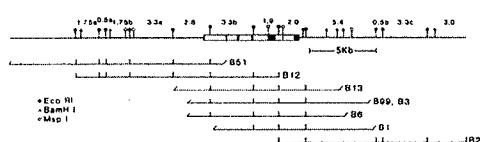
**R-Loop Analysis** - Phage clones (10  $\mu$ g/ml) containing  $\beta$ -casein DNA were hybridized to a  $\beta$ -casein mRNA-enriched Sepharose 4B column fraction (30  $\mu$ g/ml) (43) in 70% Formamide, 0.3 M NaCl, 10 mM Tris-HCl (pH 7.4), 10 mM NaEDTA, 0.2% SDS, 0.1% PMSF, 0.04% NP-40, 0.04% 2-mercaptoethanol, 0.04% polyvinylpyrrolidone, 0.04% Ficoll 400, 0.1% SDS, 0.1% NP-40, 0.04% 2-mercaptoethanol, 0.04% NaEDTA, 50 mM Tris-HCl, pH 7.4, 1.0 mM NaCl, and 300  $\mu$ g/ml sheared salmon sperm DNA (hybridization solution). After a 4 hr of hybridization at 37°C, 5  $\mu$ g/ml of lactating mammary gland poly(A)<sup>+</sup> RNA (43) was added to one of the filters and hybridization was allowed to continue for another 12 hr. The filters were washed for 10 min at 68°C in 2% SSC, 0.1% SDS and then placed in a seal-a-bag containing the hybridization solution and nick translated cloned cDNA probes ( $2.5 \times 10^6$  cpm/ml). The filters were hybridized at 37°C for 12 hr, and washed first with 2% SSC and 0.1% SDS at room temperature for 30 min, and then at 68°C for 2 hr. The filters were air dried and autoradiographed.

**RNA Analysis by In Situ Hybridization** - Nuclei were isolated at -20°C in 50% glycerol, 50 mM Tris-HCl, pH 7.5, 5 mM NaEGTA, 25 mM KCl, 0.15 mM spermidine and 0.5 mM spermine (47). The nuclear RNA was isolated by the guanidine thiocyanate/CSCl gradient method (48). Nuclear RNA was fractionated on a denaturing methyl mercury hydroxide agarose gel (49) and the  $\beta$ -casein RNA transcripts were identified by *in situ* hybridization with nick-translated cDNA probes (50).

**RESULTS**

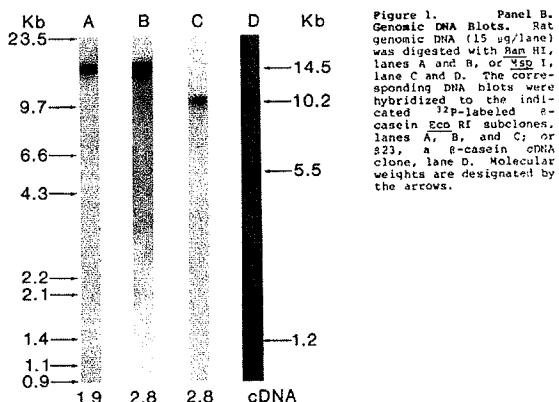
**Identification of  $\lambda$  Clones**

Two independent rat genomic DNA libraries were screened, initially with a  $\beta$ -casein cDNA clone and subsequently with  $\beta$ -casein genomic subclones. A total of 8 overlapping  $\lambda$  clones encompassing 34.4 Kb of rat DNA were isolated (Fig. 1). These clones were mapped with several restriction enzymes and the positions of the Eco RI, Bam HI and  $\lambda$  Sfi I sites are shown (Fig. 1A).



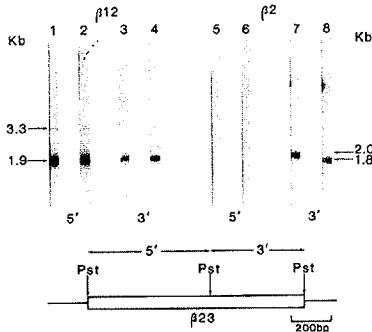
**Fig. 1.** Panel A, Map of the Rat  $\beta$ -Casein Gene Locus. Overlapping  $\beta$ -casein specific phage clones are shown in the 5' to 3' orientation. Solid boxes represent exons, open boxes represent introns, and vertical lines indicate restriction sites. Eco RI, Bam HI,  $\lambda$  Sfi I. The names of the various Eco RI subclones are indicated above the line by the numbers which correspond to their size, in Kb, in order of appearance, a, b, c, etc. The names of the various phage clones are indicated beside each clone.

To confirm that no rearrangements or gross deletions had occurred during the construction and isolation of the phage clones, a cDNA clone subcloned Eco RI genomic fragments were used as probes against restriction digests of the phage DNA (Fig. 1B). When the 1.9 Eco RI subclone, located in the middle of the gene, was used as a probe, it hybridized to only a 14.5 Kb Bam HI fragment which lies between a site within the 2.0 Eco RI fragment and a site in the 1.75b Eco RI fragment. This same Bam HI fragment is detected by the 2.8 Eco RI subclone, containing primarily 5' flanking DNA. The 2.8 Eco RI subclone detects a 10.2 Kb  $\lambda$  Sfi I fragment created by the  $\lambda$  Sfi I sites within the 1.9 Eco RI segment and a 3.3 Kb fragment within Bam 1.2 Kb  $\lambda$  Sfi I fragment. The  $\lambda$ -cDNA also detects a 5.5 Kb  $\lambda$  Sfi I fragment as expected, but the hybridization signal is expected to be less intense. The  $\lambda$ -cDNA also detects a 1.2 Kb  $\lambda$  Sfi I fragment. The latter fragment lies between the  $\lambda$  Sfi I sites of the 1.9 and 2.0 Eco RI fragments. These results indicated that the map of the genomic clones is correct and no detectable rearrangements have occurred. The absence of additional bands in the genomic DNA blots and the isolation of only one set of overlapping phage clones from two independent DNA libraries indicates that only one authentic  $\beta$ -casein gene exists in the rat genome.



## Characterization of the Genomic Clones

The  $\beta$ -casein gene was oriented and exons initially positioned within the genomic clones by DNA blots of the  $\lambda$  clones using 5' and 3'  $\beta$ -casein cDNA probes isolated from the B23 cDNA clone (Fig. 2).



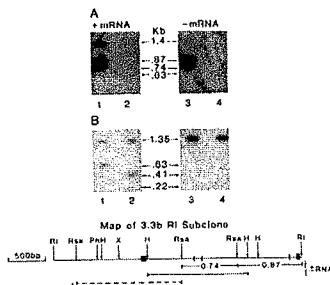
**Fig. 2.** Orientation of the  $\lambda$  Clones and Localization of Exons. Genomic phage clone B12, lanes 1-4, lanes 3-5, and 7, and Eco RI/Bam HI, lanes 2, 4, 6, 8. The DNA blots of lanes 1, 2, 5, and 6 were then hybridized with the 5' Pst I fragment of B23, shown below, which had been  $^{32}$ P-labeled. The DNA blots of lanes 3, 4, 7, and 8 were similarly hybridized with a  $^{32}$ P-labeled 3' Pst I fragment of B23. Molecular weights are indicated by the arrows.

Two genomic clones, B12 and B2, were used since they allowed the 1.9 and 2.0 Eco RI fragments to be analyzed separately. Two Eco RI fragments from the genomic B12 clone hybridized with the 5'  $\beta$ -casein cDNA probe, the 1.9 and 3.3b Kb fragments (lanes 1 and 2). The 1.9 Kb fragment also hybridized with the 3'  $\beta$ -casein cDNA probe (lanes 3 and 4). The genomic B2 clone contains only the 3' region of the  $\beta$ -casein gene since it failed to hybridize with the 5'  $\beta$ -casein cDNA probe (lanes 5 and 6). The 1.9 Kb fragment of the genomic B2 clone, which is reduced to a 1.8 Kb fragment by digestion with Eco RI/Bam HI, hybridized to the 3'  $\beta$ -casein cDNA probe indicating the presence of exon(s) within this fragment (lanes 7 and 8). The genomic B12 clone is thus 5' to the B2 clone and contains exons from both the 5' and 3' halves of the  $\beta$ -casein cDNA.

From the previous characterization of the  $\gamma$ -casein gene (13), it was known that the casein genes were large and complex genes with small exons which are difficult to detect in DNA blots by hybridization with cDNA probes. The  $\beta$ -casein cDNA probe was not able to detect a 2 Kb band in Eco RI digested genomic DNA (data not shown), which corresponds to the 2.0 Kb fragment of the B12 clone and the 1.9 Kb fragment of the B2 clone. Exons within the 3.3b Eco RI fragment were probably not detected due to their small size. In addition the B23 cDNA clone is missing the first 82 bp of the mRNA sequence and thus may not be capable of detecting some 5' exons (5). In order to detect these additional 5' exons, RNA sandwich blots were performed.

## RNA Sandwich Blots

The results of an RNA sandwich blot are shown in Fig. 3.



**Fig. 3.** RNA Sandwich Blots. The 3.3b Eco RI subclone was digested with Rsa I (lanes 1 and 2), Hinf I (lanes 3 and 4), Pst I (lanes 1, 2, 4), Hind III (lanes 1, 2, 3, 4), Kpn I (lanes 2, 4), H (lanes 1, 2, 3, 4), and R (lanes 1, 2, 3, 4). The vector bands were excised and the insert bands were bidirectionally transferred to nitrocellulose. Lanes 1 and 2, panels A and B, were then prehybridized in the presence of mammary gland lactating mRNA while lanes 3 and 4, panels A and B, were prehybridized in its absence. All lanes were then hybridized with  $^{32}$ P-labeled B23  $\beta$ -cDNA. Sizes are indicated by the arrows.

indicated by the arrows. On the map below, restriction fragments which were detected in both the presence and absence of mRNA are indicated by the solid line, while restriction fragments which were detected in only the presence of mRNA are indicated by the dashed line. Solid boxes indicate the positions of exons II-VI.

In panel A, three Rca I restriction fragments, 1.4, 0.87, and 0.74 Kb fragments of the 3.3b Eco RI subclone hybridized to the B23 cDNA probe. The 0.87 Kb and 0.74 Kb fragment, which hybridized in both the presence and absence of mRNA, must each contain at least one exon whose sequence is contained within the B23 cDNA clone. The 1.4 Kb fragment, which hybridized to the probe only in the presence of mRNA, must contain one or more exons whose sequence does not overlap the B23 cDNA clone sufficiently to allow hybridization. This exon was further localized to the 0.63 Kb fragment between the Xba I site and a Hind III site. The absence of hybridization to the 0.63 Kb Pst I/Eco RI fragment (panel A) and the hybridization to the 0.635 Kb Hind III and the 0.415 Kb Hind III/Xba I fragments (panel B). An R-loop experiment confirmed the location of these exons and located additional exons.

#### R-Loop Analysis

The genomic B12 phage clone and  $\beta$ -casein mRNA were annealed under R-loop conditions and analyzed by electron microscopy as shown in Fig. 4. From the known lengths of exons VII and intron F (see below) a scaler was determined which allowed the conversion of the observed lengths of the intron loops to Kb. By comparing the intron loop size and the  $\beta$ -casein gene map, the locations of the most 5' and 3' exon within the 3.3b Eco RI fragment, which were mapped by the RNA endonuclease blot technique described above, were confirmed. Exons VII and IV were located in the middle of the gene. The fragment failed to anneal under the R-loop conditions probably due to their small size (see below). An additional exon was found to lie 1.7 Kb 5' of exon II or approximately 500 bp from the 3' end of the 2.8 Eco RI fragment. Since restriction mapping data had located the 3' most exon to the middle of the 2.0 Eco RI fragment (data not shown) the primary transcript length of the  $\beta$ -casein gene was estimated to be approximately 7.2 Kb plus 250 bp of poly(A) for a total length of >7.5 Kb.

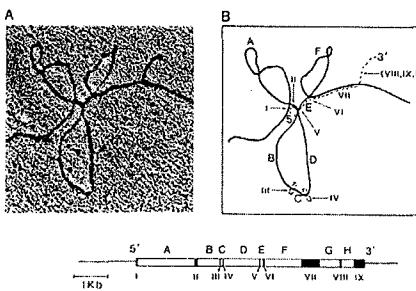


Fig. 4. R-Loop Analysis of  $\beta$ -12 Clone. B12 DNA was denatured and hybridized to Sepharose 4B fraction enriched  $\beta$ -casein mRNA as described in "Experimental Procedures". The sizes of exon VII and intron F were used to calibrate the electron micrograph in panel A and to predict the positions of exons I and II as shown in panel B.

#### RNA Blots

RNA blots using lactating poly(A)<sup>+</sup> nuclear RNA were performed in order to confirm the predicted primary transcript size. To avoid RNA degradation, the nuclei were isolated at -20°C in the presence of glycerol (46) and the RNA was isolated by a guanidine thiocyanate-CsCl protocol (47). To facilitate detection of large RNA transcripts, hybridization was performed *in situ* to gels without transfer to a filter. Using both the 1.9 Kb and the 3.3b Eco RI subclones as probes (Fig. 5), an apparent full length transcript of 7.5 Kb, identical in size to the estimated primary transcript size of 7.5 Kb was detected.

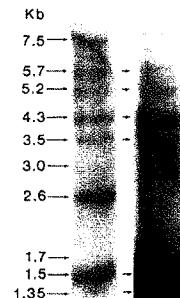


Fig. 5. Analysis of  $\beta$ -Casein Nuclear RNA Transcripts. Poly(A)<sup>+</sup> nuclear RNA from lactating mammary glands was isolated as in "Experimental Procedures". The RNA was electrophoresed on a  $\text{CH}_2\text{Cl}_2\text{OH}$  gel and hybridized *in situ* with  $^{32}\text{P}$ -labeled 1.9 (right lane) or 3.3b (left lane) Eco RI subclones (see Fig. 1).

In addition to an apparent primary  $\beta$ -casein gene transcript a number of other lower molecular weight gene products were detected (Fig. 5). Most of these were detected by both the 3.3b and 1.9 Eco RI subclone probes, suggesting that the two subclones are part of the same primary transcription unit. However, one gene product, the 2.6 Kb band, was only detected by the 3.3b probe. All true precursors must retain all the exons of the gene. Since the 2.6 Kb gene product was not detected by the 1.9 probe containing exon VII, it cannot be a true  $\beta$ -casein precursor. This gene product appears to contain all the exons of the gene except exon VII. At least two other lower molecular weight gene products were detected by the 1.9 probe. These two possible explanations can be suggested. The first is that the 3.3b probe is detecting the products of another gene due to repeats within itself. The 3.3b probe is known to contain repeats (data not shown). A second possibility is that the 2.6 Kb band is a dead end processing product of the  $\beta$ -casein gene transcripts. Whatever its origin, this 2.6 Kb RNA molecule is present in relatively high concentrations in lactating tissue.

TABLE I

EXON	SIZE	mRNA POSITION	INTRON	SIZE
I	40	1-40	A	1.7 Kb
II	63	41-103	B	0.68 Kb
III	27	104-130	C	0.12 Kb
IV	21	131-151	D	0.97 Kb
V	24	152-175	E	0.09 Kb
VI	42	176-217	F	1.1 Kb
VII	525	218-743	G	0.58 Kb
VIII	48	744-791	H	0.86 Kb
IX	326	792-1117		